Automating the Calibration of a Neonatal Condition Monitoring System

Williams, C. K. I., & Stanculescu, I. (2011). Automating the Calibration of a Neonatal Condition Monitoring System. In Peleg, M., Lavrac, N., & Combi, C. (Eds.), Artificial Intelligence in Medicine. (pp. 240-249). (Lecture Notes in Computer Science). Springer Berlin / Heidelberg. doi: 10.1007/978-3-642-22218-4_30

Automating the Calibration of a Neonatal Condition Monitoring System

Christopher K.I. Williams, Ioan Stanculescu

School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB, UK
c.k.i.williams@ed.ac.uk,i.a.stanculescu@sms.ed.ac.uk
http://www.inf.ed.ac.uk

Abstract. Condition monitoring of premature babies in intensive care can be carried out using a Factorial Switching Linear Dynamical System (FSLDS) [15]. A crucial part of training the FSLDS is the manual *calibration* stage, where an interval of normality must be identified for each baby that is monitored. In this paper we replace this manual step by using a classifier to predict whether an interval is normal or not. We show that the monitoring results obtained using automated calibration are almost as good as those using manual calibration.

Keywords: Condition monitoring, switching linear dynamical system, intensive care, logistic regression, decision tree, Naïve Bayes

1 Introduction

Condition monitoring often involves the analysis of systems with hidden factors that "switch" between different modes of operation and collectively determine the observed data. Given the monitoring data, we are interested in recovering the state of the factors that gave rise to it. In our work condition monitoring is performed on premature babies receiving intensive care, with the data coming from second-by-second measurements of their vital signs. The factors correspond to physiological events (such as bradycardia, a spontaneous slowing of the heart) or artifactual events (such as taking a blood sample).

The Factorial Switching Linear Dynamical System (FSLDS) [15] has the ability to model a system which switches between multiple modes of operation conditioned on a set of factors. More precisely, given a sequence of observations, the FSLDS outputs the filtering distribution of the switch setting at each time step. The model has proved to be highly successful in inferring the hidden factors that govern the observations collected by cotside computers [10, 12].

As a crucial part of training the FSLDS, a manual *calibration* stage is needed [12]. This requires finding an interval of normality for each examined baby. By normality, we generally understand a period in which the baby is in a stable physiological condition and there is no artifact corrupting the measurements [10]. The primary goal of this paper is *automating* the calibration stage. More precisely, we will build a binary classifier that predicts whether an interval of

monitoring data is normal or not. The main reason for the feasibility of such a classification is that while the normal dynamics can be different for each baby, artifact is stereotypical.

The structure of the rest of the paper is as follows: An introduction to physiological monitoring in a neonatal intensive care unit is given in Section 2. Section 3 is dedicated to the Factorial Switching Linear Dynamical System (FSLDS) discussing the model, the application-specific setup, learning and inference. Section 4 details our approach for automating the calibration stage needed by the model. The results obtained by employing the classifiers built in the previous part are given in Section 5. A discussion of our main findings together with recommendations for future work concludes the paper in Section 6.

2 Neonatal Condition Monitoring Data

The physiological system can be thought of in terms of three partly independent sub-systems: the respiratory system, the cardiovascular system and the thermoregulatory system [10, §2.1.1]. Each of these systems has its associated set of measurement channels. The respiratory system is monitored by measuring the O_2 saturation in arterial blood (SO) and the partial pressures of O_2 ($TcPO_2$) and CO_2 ($TcPCO_2$). The flow of blood through the body is controlled by the cardiovascular system. This is traditionally monitored by obtaining heart rate (HR) measurements from an electrocardiogram. In addition, a transducer records the evolution of the blood pressure on two channels, systolic (BS) and diastolic (BD). The thermoregulatory system keeps the body at an adequate temperature. This is monitored by two channels: core temperature (TC) and peripheral temperature (TP). Along with this set of physiological measurements, clinicians also need the environment inside the incubator to function in normal parameters. Therefore, they record incubator temperature (IT) and incubator humidity (IH).

We now enumerate the physiological and artifactual events we plan to uncover by doing inference in the auto-calibrated FSLDS. *Bradycardia* (see Figure 2.b) is a physiological event characterized by a temporary drop in the heart rate measurements. *Probe disconnection* is a frequent artifactual event related to operating the monitoring equipment. Generally, when a probe is disconnected the measurements fall to zero. However, the current paper analyses *core temperature probe detachment*, when the disconnection is characterized by a decay of measurements towards incubator values. Periodically taking a *blood sample* is another artifactual event (see Figure 2.b). The procedure causes an artifactual ramp in the blood pressure measurements. Moreover, if the heart rate is also computed from the pressure sensor, readings will cease for the duration of the blood sampling event [10, §2.3.2]. A common artifactual event is *opening the incubator's* doors. This is caused by various medical procedures that need to be performed on the patient. During this operation, we usually see an increased variance in the physiological measurement channels. At the same time the incubator's temperature and humidity slowly adjust to room values. A great number

of other factors can influence a patient's condition and precisely determining all of these is practically impossible. A solution is to introduce the *X-factor* [10], a factor responsible for all events that are neither normal nor correspond to a known factor.

3 The Factorial Switching Linear Dynamical System

Switching [3, 8, 13] and factorization [4] are two well-known ideas for relaxing the assumptions made by state-space models on the probability distribution of the data. The FSLDS [10, 12, 15] combines both with the advantages of autoregressive (AR) processes to model baby monitoring. In a traditional Switching Linear Dynamical System (SLDS) (see e.g. [13]), discrete hidden states, s_t, evolve according to Markovian transition probabilities, $p(s_t \mid s_{t-1})$, and determine which set of parameters is used by the dynamics and observation equations at the current time step:

$$\mathbf{x}_t \sim N(\mathbf{A}(s_t)\mathbf{x}_{t-1}, \mathbf{Q}(s_t)), \qquad \mathbf{y}_t \sim N(\mathbf{C}(s_t)\mathbf{x}_t, \mathbf{R}(s_t)), \tag{1}$$

where $\mathbf{x}_t$ is the continuous hidden state and $\mathbf{y}_t$ is the observed variable. The joint distribution of such a model is:

$$p(s_{1:T}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p(s_1)p(\mathbf{x}_1)p(\mathbf{y}_1 \mid \mathbf{x}_1, s_1) \prod_{t=2}^{T} p(s_t \mid s_{t-1})p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, s_t)p(\mathbf{y}_t \mid \mathbf{x}_t, s_t). \tag{2}$$

However, in problems such as physiological monitoring there are a large number

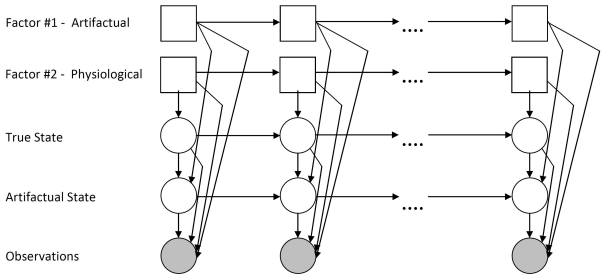


Fig. 1. The DAG of a FSLDS with 2 factors (one physiological and one artifactual). Each column represents a discrete time step. The state is divided into dimensions that approximate the "true" physiology and dimensions that approximate the artifactual patterns. Shaded variables are observed; circles represent continuous variables and squares represent discrete ones.

of factors influencing the dynamics of the system. The solution in [12] was to represent the switch variable as the cross product of M discrete factors, $s_t = f_t^{(1)} \otimes f_t^{(2)} \otimes ... \otimes f_t^{(M)}$. If the factors are assumed to be all a priori independent, the transition probabilities can be written as: $p(s_t \mid s_{t-1}) = \prod_{m=1}^{M} p(f_t^{(m)} \mid f_{t-1}^{(m)})$.

The FSLDS's structure (see Figure 1) allows valuable application-specific representational choices. For each of the measurement channels, there is precisely one visible dimension in the observation vector $\mathbf{y}_t$ at time t, and there are one or more hidden continuous dimensions in the state vector $\mathbf{x}_t$ at time t. The purpose is to increase the capability of the system to track the "true" physiological signals. Consequently, continuous hidden state dimensions can be associated either with true physiology or with artifact [12]. A similar rationale is applied to the discrete factors. However, artifactual factors can affect only artifactual states, while physiological factors can influence any state. In addition, the dynamics matrices, $\mathbf{A}(s_t)$, and dynamics noise matrices, $\mathbf{Q}(s_t)$, are chosen to have a "block diagonal" structure [12]. This is a great advantage, since the set of available observation channels usually varies from baby to baby based on the medical staff's prior beliefs about its physical condition.

Learning in the FSLDS model [10, §5] is facilitated by the fact that part of the regimes in the data are annotated by clinical experts. This means that when the hidden switch state is known, we can condition on it, making learning equivalent to training a simple Linear Dynamical System (LDS). Fortunately, there is no need to consider all possible switch settings because some factors overwrite the others [10, 12]. We can also estimate the factor transition probabilities by simple data counting.

We now discuss learning the dynamics under the normal regime, which corresponds to the LDS obtained when all the other factors are off. This stage is called *calibration* and needs to be performed separately for each baby. It requires manually selecting a period of normal measurements on all channels. The parameters are obtained by independently fitting AR processes to each of these channels [12].

Exact inference is proved to be computationally intractable in many generalizations of the state-space model [6]. Among the approximate inference methods tested so far for the FSLDS, the Gaussian sum approximation [1, 5, 8] delivered the best performance and will be exclusively used below. The basic idea is to avoid the exponential growth in the number of terms needed at each time step by applying a moment matching approximation.

Clearly the FSLDS is one of many approaches to solving the problems of condition monitoring and artifact detection. A review of this related work is beyond the scope of the present paper, but can be found in [10, §3].

4 Automating calibration

In order to automate the calibration stage, we rely on the following clinical considerations. The physiological patterns corresponding to normality are specific to each patient. On the other hand, physiological and artifactual factors like the ones introduced in Section 2 are stereotypical. This means that each occurrence of those events can be associated with a certain known pattern. In the following we describe our data, give a full problem formulation and then explain feature and classifier choices.

Exploratory Data Analysis: Our dataset consists of 24 hour recordings taken from each of fifteen premature born babies at Edinburgh Royal Infirmary. The babies were around 24 to 29 weeks gestation and aged 1 to 16 days post-partum. The data has been sampled at 1Hz and the set of measurement channels varies from baby to baby. Expert annotations are available for five known factors (Bradycardia, Blood Sample, Incubator Open, Core Temperature Probe Detachment and Transcutaneous Probe Recalibration) and for the X-factor. In addition, one period of Normal data is highlighted for each baby. These carefully chosen intervals have been used up to now to "manually" calibrate the FSLDS. As in [10, 12], due to the scarcity of examples in the dataset we will not use the Transcutaneous Probe Recalibration factor in any of the following experiments.

Counting the number of incidences of each factor, we notice that factors such as opening the incubator and the X-factor are far more frequent than taking a blood sample or a detachment of the temperature probe. We also note that the mean durations of the various events are quite different as well. Although artifactual events always respect the same patterns, there is a great deal of variability in their duration (see [14]). The consequence is that the feature extraction task becomes more challenging.

We also know that a period for which there are no annotations is automatically considered a period of normality. Thus we can compute the total duration of normality in our data, which is 283 hours (79% of the total 360 hours). Note that this computation cannot be performed by summing up the total durations of the factors, since there is a significant overlap between factors.

Problem formulation: Since the objective is to extract some periods of normality from continuously recorded data, we begin by finding an appropriate length for these intervals. Based on the duration of annotated normality periods, we choose our intervals to have a length of 15 minutes (i.e. 900 seconds). For simplicity, no overlapping is permitted. A disadvantage of fixed length intervals is that we sometimes split a single event between intervals.

Examining our annotations, we have concluded that we can use at most four known factors (Bradycardia, Blood Sample, Incubator Open and Core Temperature Probe Detachment) to assess the performance of the FSLDS. Using the clinical information summarized in Section 2, we consider the union of all the channels that are influenced by the four factors of interest: HR, BS, BD, SO, TC, IH and IT. This set is the necessary and sufficient set of channels that need to be observed in order to set up a FSLDS capable to infer the discussed factors. Our original problem of finding an interval of normality by looking at all the available channels for a baby has just reduced to looking at all the channels enumerated above. Note that this does not imply that all the channels in the set above are present for all the babies. One may also notice that introducing an observation channel not influenced by any factor in the FSLDS will have no effect on inferences because of the block diagonal structure of $\mathbf{A}(s_t)$ and $\mathbf{Q}(s_t)$.

With all this in place, we explain the "channel-based" procedure we have chosen to use for classification. We break our classification problem into seven smaller classification problems, predicting normality/non-normality for each mea-

surement channel. For these tasks, a new labelling of the data is required: If at least one of the factors has a non-empty intersection with the interval, it is labelled as being Non-Normal; otherwise it is labelled as Normal. Also note that when active, the Incubator Open and X-factor may not affect the whole set of factors they can influence, but only a subset of them. Thus, we decided to look at all available channels before making predictions for each channel.

Our reason for pursuing the "channel-based" approach is that it efficiently uses the limited amount of data on hand. An alternative "interval-based" approach making a single prediction regarding all the measurement channels would have been a poorer choice. The reason is that it is often the case that during a fifteen minute period, only a factor affecting a small subset of channels is active. This means that all the other channels are evolving normally during this period. In the "interval-based" approach this interval would have been labelled as Non-Normal, and we might have lost possibly valuable information about normality on the unaffected channels.

Feature extraction: Extracting good features is an essential requirement for success. This task is made difficult by the fact that periods of non-normailty can appear anywhere within a 15 minute interval, and that there is a significant amount of variability in the patterns of the known factors.

Normal heart rate (HR) measurements usually display a low amplitude, high frequency fluctuation around a slowly changing baseline. An event affecting this channel will generally result in a higher variance, so we chose the standard deviation as a feature. The baseline level of the heart rate signal is captured by the median feature. In order to detect bradycardia, we have chosen to record the difference between the minimum and average values of the observations. The most common event influencing blood pressure measurements (BS and BD) is taking a blood sample. The difference between the maximum and median values of these channels has been experimentally found to capture such variations. The oxygen saturation (SO) channel's dynamics can be recorded by computing the median and the difference between the median and the mean of the observations. Moving to the core temperature (TC) measurements, we are interested for these values to stay within some acceptable lower and upper limits. Thus we pick the minimum and maximum values of the channel as features. The standard deviation also offers valuable information about the baby's condition. When the incubator's doors are opened we usually see a drop in the humidity measurements (IH). Consequently, we keep track of the standard deviation of the channel and of the difference between the median and minimum values of the channel. A similar rationale is applied for the incubator temperature channel (IT).

Classifier setup: We now clarify the setting in which we have performed our experiments.

The classifiers employed for the task were logistic regression, Naïve Bayes and decision trees. These choices are mainly motivated by the simplicity, the easier interpretation of results and the reduced number of parameters associated with these classifiers. The optimization procedure used to get the maximum likelihood parameters for logistic regression is the Iterative Reweighted Least

Squares Algorithm (IRLS) [2, §4.3.3]. Our Naïve Bayes implementation models each attribute's distribution as a Gaussian. The tree building algorithm we have employed is C4.5, which relies on the information gain criterion [9].

Considering the manner in which we have chosen our (baseline) set of features, dropout measurements may raise serious problems. However, as previously stated, they can be trivially detected. Since we clearly don't want to calibrate the FSLDS using an interval that contains dropouts, we will remove periods containing such artifact from the very beginning.

Moreover, there are babies for whom we do not have all seven channels on hand. Our solution was to always input as much information as possible into our classifiers. Theoretically, we make a Missing at Random (MAR) [7] assumption about the absent measurement channels. This means we had to train separate classifiers that work on feature sets with different dimensionalities (see [14]).

5 Results

Evaluation of the auto-calibration procedure is done in two phases. First, we assess the quality of the predictions produced by the classifiers. Second, we use these predictions in order to train the FSLDS, and then run inference in the model. The latter analysis is much more interesting since it allows a direct comparison between the manual and auto-calibrated systems.

In order to avoid over-fitting, all the experiments are performed in a 3-fold cross-validation setting. For each of the three tests, ten babies are used for training and the remaining five are left for testing.

The quality of our predictions is measured by two criteria. First, we draw Receiver Operating Characteristics (ROC) curves and compute the Area Under the ROC curve (AUC), noting that the larger the better. However, our primary objective is to extract some intervals of normality from the data. This means that we do not necessarily look for the most accurate classification between Normal and Non-Normal intervals; it is sufficient for the employed classifiers to deliver some intervals that we can confidently consider to be typical for the Normal dynamics of a baby and then utilize them to calibrate the FSLDS. This consideration motivates our second criterion. We will compare the classifiers based on how well they answer the following question: *"On a per baby basis, for how many positive instances (i.e. Normal intervals) does the classifier output a posterior probability of belonging to class Normal, $P(\mathcal{C} = Normal \mid x)$, higher than the largest posterior of a negative instance (i.e. Non-Normal interval)?"*. We will call this criterion the Interval Ranking Criterion (IRC).

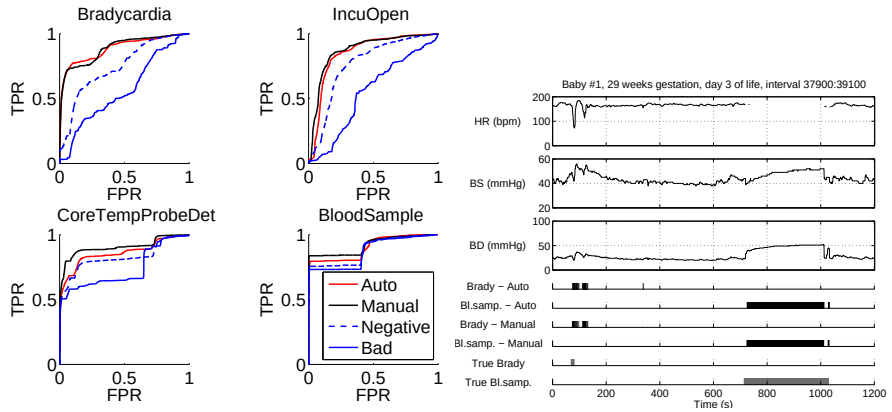
Using the baseline feature set described in the previous section, we now compare the performance of the three classifiers: logistic regression, Naïve Bayes and a decision tree. Since the features we are using display intrinsically different ranges and variances, we will standardize the input (i.e. zero mean, unit variance).

The results for the seven channel classification tasks are similar [14]. The general conclusion is that logistic regression always outperforms the other two

methods on both criteria. A distinguishing observation about logistic regression is that it has always found, for each baby, at least one positive interval with higher posterior probability of being normal than any negative instance.

In other experiments, we have studied the Bayesian approach to logistic regression, and the introduction of other features like post-natal age, gestation or even LDS parameters trained on the 15 minute intervals [14]. None of those attempts managed to outperform the classifier consisting of the baseline feature set and maximum likelihood logistic regression. We emphasize that we do not need to fix a threshold in order to use any of the classifiers above in practice. Since we just need some Normal intervals for each observed channel, we simply sort the probabilities for all the intervals corresponding to a baby and pick the top k predictions.

As previously explained, we set up a FSLDS able to infer the posterior probability distribution for four hidden factors: Incubator Open, Bradycardia, Core Temperature Probe Detachment and Blood Sample. The quality of the inferences



(a) ROC curves for the four known factors (b) Manual vs Auto calibration

Fig. 2. a) Classification results aggregated over the fifteen babies. b) Inferred distributions for Blood Sample and Bradycardia. For both methods, inference is correct. However, an inferred bradycardia instance around time $t = 125$ is in disagreement with the annotator's opinion.

will be assessed by the same two criteria as in [10, 12]. The first one is the AUC already introduced in the previous section and the second one is the equal error rate (EER)[1]. Since the EER is an error, the smaller the value the better. For evaluation, we use the same setting as the one described in [12] and all the 360

[1] The EER is the error rate computed for the threshold value at which the false positive rate (FPR) is equal to the false negative rate (FNR).

hours of physiological monitoring data on hand[2]. The experiment is again done with three-fold cross-validation: ten babies are used for training and the remaining five for testing. The auto-calibration system selects only the top prediction outputted by the classifier (i.e. $k = 1$). In Figure 2.a, we plot ROC curves aggergated over the fifteen babies corresponding to the four inferred factors for four methods of doing calibration: Auto, Manual[3] Negative and Bad. The last two are control conditions; in 'Negative' we randomly select a Non-Normal interval for calibration. In 'Bad' the we select a heavily corrupted interval for calibration. Table 1 shows that the quality of the inferences produced by the auto-calibrated

Table 1. Summary statistics for the two methods of calibration

Calibration	Statistic	Bradycardia	Incu Open	Core Temp Probe Det	Blood Sample
Auto	AUC	0.89	0.85	0.86	0.91
	EER	0.21	0.18	0.18	0.20
Manual	AUC	0.89	0.87	0.90	0.92
	EER	0.24	0.17	0.13	0.16
Negative	AUC	0.75	0.76	0.82	0.88
	EER	0.33	0.27	0.22	0.24
Bad	AUC	0.57	0.55	0.72	0.88
	EER	0.48	0.43	0.32	0.25

FSLDS is very close to the one of those produced by the manually calibrated version for three of the factors: Incubator Open, Core Temperature Probe Detachment and Blood Sample. For the remaining factor, Bradycardia, the AUC values are identical in both cases. Moreover, for this factor the auto-calibrated FSLDS manages to outperform the manual version in terms of EER. The performance deteriorates for the control conditions.

We also illustrate some comparative examples of inferences done with the manually- and auto-calibrated FSLDSs for physiological condition monitoring in Figure 2.b. The horizontal bars in the lower part of the figures indicate the posterior distributions of factors. Levels of grey from white to black indicate values from zero to one respectively. We observe that the two systems perform equally well at inferring Bradycardia and Blood Sample.

6 Discussion

In this paper we have introduced a classification-based approach to determining the normality/non-normality of intervals of monitoring data. Using carefully chosen features and a logistic regression classifier, we have demonstrated that the manual calibration stage used by the FSLDS for neonatal condition monitoring can be replaced by an automated procedure, with very little loss of performance.

[2] The experiments made use of John Quinn's code for the FSLDS [11].

[3] The results obtained with the manually-calibrated FSLDS are not identical to the ones in [12] due to using an updated version of both code and data annotations.

This reduction in the need for manual input (and consequent error) should be of great benefit in the clinical context.

The work on auto-calibration can be extended in a number of directions. We can consider alternatives to the fixed-length no-overlapping constraint imposed on the intervals used for prediction. In terms of evaluation, we can use an event-based detection analysis, as opposed to the current second-by-second inference.

Acknowledgments. We thank John Quinn and the staff of the Simpson Centre for Reproductive Health, Royal Infirmary of Edinburgh for their assistance with this work. This work is supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

References

1. Daniel L. Alspach and Harold W. Sorenson. Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.
2. Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.
3. Zoubin Ghahramani and Geoffrey E. Hinton. Variational Learning for Switching State-Space Models. *Neural Computation*, 12(4):831–864, 2000.
4. Zoubin Ghahramani and Michael I. Jordan. Factorial Hidden Markov Models. *Machine Learning*, 29:245–273, 1997.
5. C-J. Kim. Dynamic Linear Models with Markov-Switching. *J.Econometrics*, 60:1–22, 1994.
6. Uri Lerner and Ronald Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In *UAI*, pages 310–318, 2001.
7. R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. New York, Wiley, 1987.
8. Kevin P. Murphy. Switching Kalman Filters. Technical report, U.C. Berkeley, 1998.
9. J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
10. John Quinn. *Bayesian Condition Monitoring in Neonatal Intensive Care*. PhD thesis, University of Edinburgh, 2007. http://hdl.handle.net/1842/2144.
11. John Quinn. Neonatal condition monitoring demonstration code, 2008. http://omnipresence.org/jq/software.html.
12. John A. Quinn, Christopher K. I. Williams, and Neil McIntosh. Factorial Switching Linear Dynamical Systems Applied to Physiological Condition Monitoring. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(9):1537–1551, 2009.
13. R. Shumway and D. Stoffer. Dynamic linear models with switching. *J. of the American Statistical Association*, 86:763–769, 1991.
14. Ioan Stanculescu. Auto-Calibration for Neonatal Condition Monitoring. Master's thesis, University of Edinburgh, School of Informatics, 2010.
15. Christopher K. I. Williams, John A. Quinn, and Neil McIntosh. Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care. In Y Weiss, B Schölkopf, and J Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, 2006.